UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/662,985	09/15/2003	Min Chu	M61.12-0565	2246

27366 7590 09/05/2008

WESTMAN CHAMPLIN (MICROSOFT CORPORATION)
SUITE 1400
900 SECOND AVENUE SOUTH
MINNEAPOLIS, MN 55402-3244

EXAMINER	
COLUCCI, MICHAEL C	

ART UNIT	PAPER NUMBER
2626	

MAIL DATE	DELIVERY MODE
09/05/2008	PAPER

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Office Action Summary	Application No.	Applicant(s)	
	10/662,985	CHU ET AL.	
	Examiner	Art Unit	
	MICHAEL C. COLUCCI	2626	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☐ Responsive to communication(s) filed on _____.

2a)☒ This action is **FINAL**. 2b)☐ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) *23,25,26,28,29,31 and 32* is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) *23,25,26,28,29,31 and 32* is/are rejected.

7)☐ Claim(s) _____ is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☐ The specification is objected to by the Examiner.

10)☒ The drawing(s) filed on *15 September 2003* is/are: a)☒ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)☒ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☒ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☐ Notice of References Cited (PTO-892)
2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)
3) ☒ Information Disclosure Statement(s) (PTO/SB/08) Paper No(s)/Mail Date _____.
4) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____.
5) ☐ Notice of Informal Patent Application
6) ☐ Other: _____.

DETAILED ACTION

Response to Arguments

1. Applicant's arguments filed 06/11/2008 have been fully considered but they are not persuasive.

Argument 1 (page 9 paragraph 1):

- "Applicants respectfully fail to see how the cited portion of Seide teaches, suggests or renders obvious "identifying a set of candidate speech segments for a speech unit comprises applying the context information for a speech unit to a decision tree to identify a leaf node containing candidate speech segments for the speech unit" as recited in amended claim 23. Although Seide describes a tree having nodes, the detailed discussion does not pertain to speech synthesis at all, much less applying context information as presently recited."

Response to argument 1:

Examiner takes the position that the primary reference Huang et al "Recent improvements on Microsoft's trainable text-to- speech system-Whistler" (hereinafter Huang) specifically teaches speech synthesis that in order to achieve a more natural voice quality, one must take more contexts into account, going beyond diphones. However, simply modeling triphones (a phone with a specific left and right context) already requires more than 10,000 units for English. Fortunately, effective clustering of similar contexts modeled in a sub-phonetic level, to allow flexible memory-quality compromise, has been well studied in the

speech recognition community [6]. Whistler uses decision tree based senones [3][8] as the synthesis units. A senone is a context-dependent subphonetic unit which is equivalent to a HMM state in a triphone (which can be easily extended to more detailed context dependent phones, like quinphone). A senone could represent an entire triphone if a 1-state HMM is used to model each phoneme. The senone decision trees are generated automatically from the analysis database to obtain minimum within-unit distortion (or entropy). As widely used in speech recognition, the use of decision trees will generalize to contexts not seen in the training data based on phonetic categories of neighboring contexts, yet will provide detailed models for contexts that are represented in the database (page 961 Col. 2 paragraph 1 & section 3.2).

Further, Huang teaches speech synthesis and contextual analysis to match the best speech units for a synthesizer, wherein the output will be more natural with enhanced quality relative to a statistical approach and prosodic modeling. Huang teaches that we first compute unit statistics for amplitude, pitch and duration, and remove those instances far away from the unit mean. Of the remaining unit instances, a small number can be selected through the use of an objective function. In our current implementation, the objective function is based on HMM scores. During runtime, the synthesizer could either concatenate the best units preselected in the off-line analysis or dynamically select the senone instance sequence that minimizes a joint distortion function. The joint distortion function is

a combination of HMM score, unit concatenation distortion and prosody mismatch distortion. Experiments indicate that our multiple instance based synthesizer significantly improve the naturalness and overall quality over traditional single instance diphone synthesizer because of its rich context modeling, including phonetic, spectral and prosodic contexts. (Huang page 961 Col. 2 paragraph 3).

Furthermore, the secondary reference of Seide US 5857169 A (hereinafter Seide) was introduced to strongly cover the use of leaf-node relationships relevant to speech/text analysis, wherein grammar, speech features, syntactical analysis, and unit matching are enforced in direct relationship to speech segments. Seide, like Huang, teaches well known methods that utilize Markov modeling and speech recognition to acquire samples of speech. Once the samples are obtained, the analysis takes place wherein Seide teaches locating elements within the speech (in the form of vectors). Seide teaches a localizer 50 performs the locating by, for each observation vector, searching the tree structure corresponding to a reference unit until at the lowest tree level a number of leaf nodes are selected. For the selected leaf nodes, the localizer 50 determines how well the observation vector matches this reference unit. This involves for each selected leaf node using the reference probability density, which corresponds to the leaf node, to calculate an observation likelihood for the observation vector. For each reference unit, the observation likelihoods, which

have been calculated for one observation vector, are combined to give a reference unit similarity score. For each reference pattern, the reference unit similarity scores of the reference unit, which correspond to the reference pattern are combined to form a pattern similarity score. This is repeated for successive observation vectors. The reference pattern for which an optimum, such as a maximum likelihood, is calculated for the pattern similarity score is located as the recognized pattern. The description focuses on locating reference probability densities and calculating observation likelihoods. It is well understood in the art how this key element can be used in combination with other techniques, such as Hidden Markov Models, to recognize a time sequential pattern, which is derived from a continual physical quantity. It is also well understood in the art how techniques, such as a leveled approach, can be used to recognize patterns which comprise a larger sequence of observation vectors than the reference patterns. For instance, it is known how to use sub-word units as reference patterns to recognize entire words or sentences. It is also well understood how additional constraints, such as a pronunciation lexicon and grammar, may be placed on the pattern recognition. The additional information, such as the pronunciation lexicon, can be stored using the same memory as used for storing the reference pattern database (Seide Col. 8 lines 31-67).

Therefore, by combining the references of Huang and Seide, an even stronger and tangible method of speech analysis is produced, wherein Seide strengthens

Huang's contextual-based speech unit analysis through the use of various applied probability distributions, leaf-node ranking, and storage routines, wherein the overall speech synthesizer of Huang in view of Seide allow for an optimized output of natural sounding speech based on prosodic, lexical, and syntactical features as well as grammatical analysis to produce the highest matching score (Seide Col. 8 lines 31-67).

Claim Rejections - 35 USC § 103

2. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

> (a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

3. Claims 23, 25, 26, 28, 29, 31, and 32 are rejected under 35 U.S.C. 103(a) as being unpatentable over Huang et al "Recent improvements on Microsoft's trainable text-to-speech system-Whistler" (hereinafter Huang) in view of Seide US 5857169 A (hereinafter Seide).

Re claims 23 and 28, Huang teaches a method of selecting speech segments for concatenative speech synthesis (page 961 Col. 1 3.1 Unit Generation) the method comprising:

parsing an input text into speech units (page 961 Col. 2 paragraph 1-2);

identifying context information for each speech unit based on its location in the input text and at least one neighboring speech unit (page 961 Col. 2 paragraph 1);

identifying a set of candidate speech segments for each speech unit based on the context information (page 960 Col. 2 *Prosody contour generation*), wherein identifying a set of candidate speech segments for a speech unit comprises applying the context information for a speech unit to a decision tree (page 961 Col. 2 paragraph 1), wherein identifying a sequence of speech segments comprises using an objective measure comprising one or more first order components from a set of factors comprising:

an indication of a position of a speech unit in a phrase;

an indication of a position of a speech unit in a word;

an indication of a category for a phoneme preceding a speech unit (page 961 Col. 2 paragraph 1);

an indication of a category for a phoneme following a speech unit (page 961 Col. 2 paragraph 1);

an indication of a category for tonal identity of the current speech unit;

an indication of a category for tonal identity of a preceding speech unit;

an indication of a category for tonal identity of a following speech unit;

an indication of a level of stress of a speech unit;

an indication of a coupling degree of pitch, duration and/or energy with a neighboring unit;

an indication of a degree of spectral mismatch with a neighboring speech unit.

identifying a sequence of speech segments from the candidate speech segments (page 960 Col. 2 *Prosody contour generation*) based in part on a smoothness cost between the speech segments (page 960 Col. 2 *Stochastic variation* & *Contour Interpolation and Smoothing*); and

generating synthesized speech using the sequence of speech segments without further prosody modification (page 960 section 2.2 Prosody Model)

However Huang fails to particularly teach to identify a leaf node containing candidate speech segments for the speech unit

Seide teaches a localizer 50 performs the locating by, for each observation vector, searching the tree structure corresponding to a reference unit until at the lowest tree level a number of leaf nodes are selected. For the selected leaf nodes, the localizer 50 determines how well the observation vector matches this reference unit. This involves for each selected leaf node using the reference probability density, which corresponds to the leaf node, to calculate an observation likelihood for the observation vector. For each reference unit, the observation likelihoods, which have been calculated for one observation vector, are combined to give a reference unit similarity score. For each reference pattern, the reference unit similarity scores of the reference unit, which correspond to the reference pattern are combined to form a pattern similarity score. This is repeated for successive observation vectors. The reference pattern for which an optimum, such as a maximum likelihood, is calculated for the pattern similarity score is located as the recognized pattern. The description focuses on locating

reference probability densities and calculating observation likelihoods. It is well understood in the art how this key element can be used in combination with other techniques, such as Hidden Markov Models, to recognize a time sequential pattern, which is derived from a continual physical quantity. It is also well understood in the art how techniques, such as a leveled approach, can be used to recognize patterns which comprise a larger sequence of observation vectors than the reference patterns. For instance, it is known how to use sub-word units as reference patterns to recognize entire words or sentences. It is also well understood how additional constraints, such as a pronunciation lexicon and grammar, may be placed on the pattern recognition. The additional information, such as the pronunciation lexicon, can be stored using the same memory as used for storing the reference pattern database (Seide Col. 8 lines 31-67).

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention to modify the system of Huang to incorporate identifying a leaf node containing candidate speech segments for the speech unit as taught by Seide to allow for an optimized output of natural sounding speech based on prosodic, lexical, and syntactical features as well as grammatical analysis to produce the highest matching score (Seide Col. 8 lines 31-67).

Re claims 25, 29, and 31, Huang teaches the method of claim 23 wherein identifying a set of candidate speech segments further comprises pruning some speech segments from a leaf node based on differences between the context information of the

speech unit from the input text and context information associated with the speech segment (page 961 Col. 2 paragraph 3)

However, Huang fails to teach pruning some speech segments from a leaf node

Seide teaches a localizer 50 performs the locating by, for each observation vector, searching the tree structure corresponding to a reference unit until at the lowest tree level a number of leaf nodes are selected. For the selected leaf nodes, the localizer 50 determines how well the observation vector matches this reference unit. This involves for each selected leaf node using the reference probability density, which corresponds to the leaf node, to calculate an observation likelihood for the observation vector. For each reference unit, the observation likelihoods, which have been calculated for one observation vector, are combined to give a reference unit similarity score. For each reference pattern, the reference unit similarity scores of the reference unit, which correspond to the reference pattern are combined to form a pattern similarity score. This is repeated for successive observation vectors. The reference pattern for which an optimum, such as a maximum likelihood, is calculated for the pattern similarity score is located as the recognized pattern. The description focuses on locating reference probability densities and calculating observation likelihoods. It is well understood in the art how this key element can be used in combination with other techniques, such as Hidden Markov Models, to recognize a time sequential pattern, which is derived from a continual physical quantity. It is also well understood in the art how techniques, such as a leveled approach, can be used to recognize patterns which comprise a larger sequence of observation vectors than the reference patterns. For

instance, it is known how to use sub-word units as reference patterns to recognize entire words or sentences. It is also well understood how additional constraints, such as a pronunciation lexicon and grammar, may be placed on the pattern recognition. The additional information, such as the pronunciation lexicon, can be stored using the same memory as used for storing the reference pattern database (Seide Col. 8 lines 31-67).

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention to modify the system of Huang to incorporate pruning some speech segments from a leaf node as taught by Seide to allow for an optimized output of natural sounding speech based on prosodic, lexical, and syntactical features as well as grammatical analysis to produce the highest matching score (Seide Col. 8 lines 31-67).

Re claims 26 and 32, Huang teaches the method of claim 23 wherein identifying a sequence of speech segments comprises using a smoothness cost (page 960 Col. 2 *Stochastic variation* & *Contour Interpolation and Smoothing*) that is based on whether two neighboring candidate speech segments appeared next to each other in a training corpus (page 961 Col. 2 paragraph 1).

NOTE: For purposes of prior art a smoothness cost is construed to be functionally equivalent and effective as a contour interpolation an smoothing component, used to create a more natural effect. Huang teaches a natural speech synthesis, where if an exact match is not found, a cost function is employed to synthesis the most natural speech.

Conclusion

4. **THIS ACTION IS MADE FINAL.** Applicant is reminded of the extension of time policy as set forth in 37 CFR 1.136(a).

A shortened statutory period for reply to this final action is set to expire THREE MONTHS from the mailing date of this action. In the event a first reply is filed within TWO MONTHS of the mailing date of this final action and the advisory action is not mailed until after the end of the THREE-MONTH shortened statutory period, then the shortened statutory period will expire on the date the advisory action is mailed, and any extension fee pursuant to 37 CFR 1.136(a) will be calculated from the mailing date of the advisory action. In no event, however, will the statutory period for reply expire later than SIX MONTHS from the mailing date of this final action.

Any inquiry concerning this communication or earlier communications from the examiner should be directed to Michael C. Colucci whose telephone number is (571)-270-1847. The examiner can normally be reached on 9:30 am - 6:00 pm, Monday-Friday.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Richemond Dorvil can be reached on (571)-272-7602. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

/Michael C Colucci/
Examiner, Art Unit 2626
Patent Examiner
AU 2626
(571)-270-1847
Michael.Colucci@uspto.gov

/Richemond Dorvil/
Supervisory Patent Examiner, Art Unit 2626